A Unified Closed-Loop Stability Measure for Finite-Precision Digital Controller Realizations Implemented in Different Representation Schemes

Jun Wu, Sheng Chen, James F. Whidborne, and Jian Chu

***Abstract*—A computationally tractable unified finite word length closed-loop stability measure is derived which is applicable to fixed-point, floating-point and block-floating-point representation schemes. Both the dynamic range and precision of an arithmetic scheme are considered in this new unified measure. For each arithmetic scheme, the optimal controller realization problem is defined and a numerical optimization approach is adopted to solve it. Numerical examples are used to illustrate the design procedure and to compare the optimal controller realizations in different representation schemes.**

***Index Terms*—Closed-loop stability, digital controller, finite word length, number representation format, optimization.**

I. Introduction

In recent years, there has been a growing interest in digital controller implementation which reduces the finite word length (FWL) effects on closed-loop stability (see [1], [2], and the references therein). It is well known that a control law can be accomplished with different realizations and that the parameters of a controller realization are represented by a digital processor of finite bit length in a particular format, namely fixed-point, floating-point, or block-float-point format. Previous works [3]–[8] have derived various FWL closed-loop stability measures for these three formats separately and defined corresponding optimal controller realization problems based on these measures. However, all these previous measures are only linked to the precision bits of the respective representation schemes used and they do not consider the dynamic range bits. Arguably, a better approach is to consider some measure which has a direct link to the total bit length required. The main contribution of this note is to derive a unified FWL closed-loop stability measure that can accommodate both the dynamic range and precision requirements and is applicable to all the three schemes.

II. Representation Schemes

The fixed-point format with a bit length $\beta = 1+\beta_g+\beta_f$ represents a real number $x \in \mathcal{R}$ by assigning 1 bit for the sign, β_g bits for the integer part, and β_f bits for the fraction part of x. Assuming no overflow, which means that $|x| \leq 2^{\beta_g}$, x is perturbed to

$$\mathcal{Q}_1(x) = x + \delta_1 \quad |\delta_1| < 2^{-(\beta_f+1)}. \tag{1}$$

Any $x \in \mathcal{R}$ can be expressed uniquely as $x = (-1)^s \times w \times 2^e$, where $s \in \{0,1\}$ is the sign of x, $w \in [0.5, 1)$ is the mantissa of x, $e = \lfloor \log_2 |x| \rfloor + 1 \in \mathcal{Z}$ is the exponent of x, $\mathcal{Z}$ denotes the set of integers and the *floor* function $\lfloor x \rfloor$ is the closest integer less than or equal to x. The floating-point format with a bit length $\beta = 1+\beta_w+\beta_e$ represents x by assigning 1 bit for s, β_w bits for w and β_e bits for e. Let $\underline{e}$ and $\overline{e}$ be the lower and upper limits of the exponent, respectively. Clearly, $\overline{e} - \underline{e} = 2^{\beta_e} - 1$. Denote the set of integers $\underline{e} \leq e \leq \overline{e}$ as $\mathcal{Z}_{[\underline{e},\overline{e}]}$. Assuming that no underflow or overflow occurs, which means that the exponent of x is within $\mathcal{Z}_{[\underline{e},\overline{e}]}$, x is perturbed to [7]

$$\mathcal{Q}_2(x) = x + x\delta_2 \quad |\delta_2| < 2^{-(\beta_w+1)}. \tag{2}$$

In the block-floating-point format, a set of real numbers $\mathcal{S}$ is first divided into some blocks. For an illustrative purpose, consider the case of dividing $\mathcal{S}$ into the two nonempty and nonoverlapped subsets $\mathcal{S}_1$ and $\mathcal{S}_2$. Let $\eta_1 \in \mathcal{S}_1$ be the element in $\mathcal{S}_1$ that has the largest absolute value, and $\eta_2 \in \mathcal{S}_2$ be the element in $\mathcal{S}_2$ that has the largest absolute value. Then, any $x \in \mathcal{S}$ can be expressed uniquely as $x = (-1)^s \times u \times 2^h$, where $u \in [0,1)$ is the block mantissa of x, and the block exponent of x is

$$h \triangleq \begin{cases} \lfloor \log_2 |\eta_1| \rfloor + 1, & \text{for } x \in S_1 \\ \lfloor \log_2 |\eta_2| \rfloor + 1, & \text{for } x \in S_2 \end{cases}. \tag{3}$$

When all the elements in $\mathcal{S}$ are presented in the block-floating-point format of bit length $\beta = 1+\beta_u+\beta_h$, the bits are assigned as follows: 1 bit for the sign, β_u bits for u which is represented in fixed-point with the two's complement system, and β_h bits for h. Let $\underline{h}$ and $\overline{h}$ be the lower and upper limits of the block exponent, respectively. Obviously, $\overline{h} - \underline{h} = 2^{\beta_h} - 1$. Denote

$$r(x) \triangleq \begin{cases} 2\eta_1, & \text{for } x \in S_1 \\ 2\eta_2, & \text{for } x \in S_2 \end{cases}. \tag{4}$$

Assuming no underflow or overflow, i.e., the block exponent of x is within $\mathcal{Z}_{[\underline{h},\overline{h}]}$, it can be shown that x is perturbed to

$$\mathcal{Q}_3(x) = x + r(x)\delta_3 \quad |\delta_3| < 2^{-(\beta_u+1)}. \tag{5}$$

It is easily seen that in each representation format the total bit length always consists of three parts. Sign occupies one bit. The dynamic range of representation is defined by β_g, β_e, or β_h bits, and the precision of representation is determined by β_f, β_w, or β_u bits, depending on which scheme is actually chosen. For notational conciseness, we introduce the "generalized" dynamic range bit length β_r and precision bit length β_p for the three representation schemes. It is understood that $\beta_r = \beta_g$, β_e, or β_h and $\beta_p = \beta_f$, β_w or β_u, depending on which format is actually used.

III. Problem Statement

The discrete-time linear time-invariant plant P is described by

$$\begin{cases} \mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \mathbf{B}\mathbf{e}(k) \\ \mathbf{y}(k) = \mathbf{C}\mathbf{x}(k) \end{cases} \tag{6}$$

with $\mathbf{A} \in \mathcal{R}^{n\times n}$, $\mathbf{B} \in \mathcal{R}^{n\times p}$, and $\mathbf{C} \in \mathcal{R}^{q\times n}$. The generic digital controller C is described by

$$\begin{cases} \mathbf{v}(k+1) = \mathbf{F}\mathbf{v}(k) + \mathbf{G}\mathbf{y}(k) + \mathbf{H}\mathbf{e}(k) \\ \mathbf{u}(k) = \mathbf{J}\mathbf{v}(k) + \mathbf{M}\mathbf{y}(k) \end{cases} \tag{7}$$

with $\mathbf{F} \in \mathcal{R}^{m\times m}$, $\mathbf{G} \in \mathcal{R}^{m\times q}$, $\mathbf{J} \in \mathcal{R}^{p\times m}$, $\mathbf{M} \in \mathcal{R}^{p\times q}$, and $\mathbf{H} \in \mathcal{R}^{m\times p}$. Let $\mathbf{e}(k) = \mathbf{q}(k) + \mathbf{u}(k)$ with the command input $\mathbf{q}(k)$. Then, P and C form a discrete-time closed-loop control system.

Assume that a realization ($\mathbf{F}_0$, $\mathbf{G}_0$, $\mathbf{J}_0$, $\mathbf{M}_0$, $\mathbf{H}_0$) of C has been designed. It is well-known that the realizations of C are not unique. All the realizations of C form the realization set

$$Sc \triangleq \{(\mathbf{F},\mathbf{G},\mathbf{J},\mathbf{M},\mathbf{H}) : \mathbf{F} = \mathbf{T}^{-1}\mathbf{F}_0\mathbf{T}, \mathbf{G} = \mathbf{T}^{-1}\mathbf{G}_0 \\ \mathbf{J} = \mathbf{J}_0\mathbf{T}, \mathbf{M} = \mathbf{M}_0, \mathbf{H} = \mathbf{T}^{-1}\mathbf{H}_0\} \tag{8}$$

where $\mathbf{T} \in \mathcal{R}^{m\times m}$ is any real-valued nonsingular matrix, called a similarity transformation. Let $\mathbf{w}_F \triangleq \mathrm{Vec}(\mathbf{F})$, where $\mathrm{Vec}(\cdot)$ denotes the

Manuscript received January 18, 2002; revised September 30, 2002. Recommended by Associate Editor D. E. Miller. The work of J. Wu and S. Chen was supported by the U.K. Royal Society under a KC Wong Fellowship (RL/ART/CN/XFI/KCW/11949). The work of J. Wu and J. Chu was supported by the National Natural Science Foundation of China under Grant 60174026.

J. Wu and J. Chu are with the National Key Laboratory of Industrial Control Technology Institute of Advanced Process Control Zhejiang University, Hangzhou 310027, China.

S. Chen is with the Department of Electronics and Computer Science University of Southampton, Highfield, SO17 1BJ Southampton, U.K.

J. F. Whidborne is with the Department of Mechanical Engineering King's College London, Strand, WC2R 2LS London, U.K.

Digital Object Identifier 10.1109/TAC.2003.811260

column stacking operator. The vectors $\mathbf{w}_{F_0}, \mathbf{w}_G, \mathbf{w}_{G_0}, \mathbf{w}_J, \mathbf{w}_{J_0}, \mathbf{w}_M, \mathbf{w}_{M_0}, \mathbf{w}_H$, and $\mathbf{w}_{H_0}$ are similarly defined. Denote

$$\mathbf{w} = [w_1 \cdots w_N]^T \triangleq \left[\mathbf{w}_F^T \mathbf{w}_G^T \mathbf{w}_J^T \mathbf{w}_M^T \mathbf{w}_H^T\right]^T, \quad \mathbf{w}_0 \triangleq \left[\mathbf{w}_{F_0}^T \mathbf{w}_{G_0}^T \mathbf{w}_{J_0}^T \mathbf{w}_{M_0}^T \mathbf{w}_{H_0}^T\right] \tag{9}$$

where $N = (m+p)(m+q)+mp$ and ^{T} is the transpose operator. We also refer to $\mathbf{w}$ as a realization of C. The stability of the closed-loop system depends on the eigenvalues of the matrix

$$\overline{\mathbf{A}}(\mathbf{w}) \triangleq \begin{bmatrix} \mathbf{A}+\mathbf{BMC} & \mathbf{BJ} \\ \mathbf{GC}+\mathbf{HMC} & \mathbf{F}+\mathbf{HJ} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & 0 \\ 0 & \mathbf{T}^{-1} \end{bmatrix} \overline{\mathbf{A}}(\mathbf{w}_0) \begin{bmatrix} \mathbf{I} & 0 \\ 0 & \mathbf{T} \end{bmatrix} \tag{10}$$

where $\mathbf{0}$ and $\mathbf{I}$ denote the zero and identity matrices of appropriate dimensions, respectively. All the different realizations $\mathbf{w}$ have the same set of closed-loop poles if they are implemented with infinite precision. Since the closed-loop system is designed to be stable, the eigenvalues $|\lambda_i(\overline{\mathbf{A}}(\mathbf{w}))| = |\lambda_i(\overline{\mathbf{A}}(\mathbf{w}_0))| < 1, \forall i \in \{1, \ldots, m+n\}$.

Define the index α of representation formats

$$\alpha = \begin{cases} 1, & \text{fixed-point format is adopted} \\ 2, & \text{floating-point format is adopted} \\ 3, & \text{block-floating-point format is adopted} \end{cases}. \tag{11}$$

The controller realization $\mathbf{w}$ is implemented in format α of β_r dynamic range bits, β_p precision bits and one sign bit. In the remainder of this note, it is assumed that if $\mathbf{w}$ is stored in the block-floating-point format, it is divided into "natural" blocks of $\mathbf{w}_F$, $\mathbf{w}_G$, $\mathbf{w}_J$, $\mathbf{w}_M$ and $\mathbf{w}_H$. Let $\eta_F \in \mathbf{w}_F$ be the element in $\mathbf{F}$ which has the largest absolute value. The elements η_G, η_J, η_M and η_H are similarly defined. Denote

$$\begin{aligned} \|\mathbf{w}\|_{\max} &\triangleq \max_{j\in\{1,\cdots,N\}} |w_j| \\ \pi(\mathbf{w}) &\triangleq \min_{j\in\{1,\cdots,N\}} \{|w_j| : w_j \neq 0\} \\ \mathbf{z}(\mathbf{w}) &\triangleq [\eta_F \ \eta_G \ \eta_J \ \eta_M \ \eta_H]^T. \end{aligned} \tag{12}$$

Firstly, the dynamic range of β_r bits must be large enough for $\mathbf{w}$. We define a dynamic range measure for controller realization $\mathbf{w}$ in format α as

$$\gamma(\mathbf{w},\alpha) \triangleq \begin{cases} \|\mathbf{w}\|_{\max}, & \alpha = 1 \\ \log_2 \frac{4\|\mathbf{w}\|_{\max}}{\pi(\mathbf{w})}, & \alpha = 2 \\ \log_2 \frac{4\|\mathbf{z}(\mathbf{w})\|_{\max}}{\pi(\mathbf{z}(\mathbf{w}))}, & \alpha = 3 \end{cases}. \tag{13}$$

Proposition 1: The realization $\mathbf{w}$ can be represented in format α of β_r dynamic range bits without overflow ($\alpha = 1$) or without underflow or overflow ($\alpha = 2, 3$), if $2^{\beta_r} \geq \gamma(\mathbf{w},\alpha)$.

Proof: The proof is straightforward. Here, we only give the case of $\alpha = 3$. When $2^{\beta_h} \geq \log_2(\|\mathbf{z}(\mathbf{w})\|_{\max}/\pi(\mathbf{z}(\mathbf{w}))) + 2$, we have

$$2^{\beta_h} - 1 \geq \log_2\left(\frac{\|\mathbf{z}(\mathbf{w})\|_{\max}}{\pi(\mathbf{z}(\mathbf{w}))}\right) + 1 \geq \left(\lfloor \log_2 \|\mathbf{z}(\mathbf{w})\|_{\max} \rfloor + 1\right) - \left(\lfloor \log_2 \pi(\mathbf{z}(\mathbf{w})) \rfloor + 1\right). \tag{14}$$

According to the results of Section II, this means that η_F, η_G, η_J, η_M and η_H can all be represented without underflow or overflow and, therefore, $\mathbf{w}$ can be represented in the block-floating-point format of β_h block exponent bits without underflow or overflow.

Let $\beta_r^{\min}$ be the smallest dynamic-range bit length that, when used to implement $\mathbf{w}$, does not cause overflow or underflow. This minimum dynamic-range bit length can easily be computed by

$$\beta_r^{\min}(\mathbf{w},\alpha) = \begin{cases} \lceil \log_2 \|\mathbf{w}\|_{\max} \rceil, & \alpha = 1 \\ \lceil \log_2(\lfloor \log_2 \|\mathbf{w}\|_{\max} \rfloor - \lfloor \log_2 \pi(\mathbf{w}) \rfloor + 1) \rceil, & \alpha = 2 \\ \lceil \log_2(\lfloor \log_2 \|\mathbf{z}(\mathbf{w})\|_{\max} \rfloor - \lfloor \log_2 \pi(\mathbf{z}(\mathbf{w})) \rfloor + 1) \rceil, & \alpha = 3 \end{cases} \tag{15}$$

where the *ceiling* function $\lceil x \rceil$ denotes the closest integer greater than or equal to $x \in \mathcal{R}$. Note that the measure $\gamma(\mathbf{w},\alpha)$ defined in (13) provides an estimate of $\beta_r^{\min}$ as $\hat{\beta}_r^{\min}(\mathbf{w},\alpha) \triangleq \lceil \log_2 \gamma(\mathbf{w},\alpha) \rceil$. It can easily be seen that $\hat{\beta}_r^{\min} \geq \beta_r^{\min}$.

For a vector $\mathbf{x}$, let $\mathbf{d}(\mathbf{x})$ be the vector of the same dimension whose elements are all 1's, and denote

$$\tau(\mathbf{x}) \triangleq \begin{cases} 0, & \mathbf{x} \text{ is a zero vector} \\ 1, & \mathbf{x} \text{ is a nonzero vector} \end{cases}. \tag{16}$$

For two vectors $\mathbf{x} = [x_j]$ and $\mathbf{y} = [y_j]$ of the same dimension, define the Hadamard product of $\mathbf{x}$ and $\mathbf{y}$ as $\mathbf{x} \circ \mathbf{y} \triangleq [x_j y_j]$. When the dynamic range is sufficient, according to the results of Section II, $\mathbf{w}$ is perturbed to $\mathbf{w} + \mathbf{r}(\mathbf{w},\alpha) \circ \mathbf{\Delta}$ due to finite β_p where

$$\begin{aligned} \mathbf{r}(\mathbf{w},1) &= \begin{bmatrix} \tau(\mathbf{w}_F)\mathbf{d}(\mathbf{w}_F) \\ \tau(\mathbf{w}_G)\mathbf{d}(\mathbf{w}_G) \\ \tau(\mathbf{w}_J)\mathbf{d}(\mathbf{w}_J) \\ \tau(\mathbf{w}_M)\mathbf{d}(\mathbf{w}_M) \\ \tau(\mathbf{w}_H)\mathbf{d}(\mathbf{w}_H) \end{bmatrix} \\ \mathbf{r}(\mathbf{w},2) &= \mathbf{w} \\ \mathbf{r}(\mathbf{w},3) &= \begin{bmatrix} 2\eta_F \mathbf{d}(\mathbf{w}_F) \\ 2\eta_G \mathbf{d}(\mathbf{w}_G) \\ 2\eta_J \mathbf{d}(\mathbf{w}_J) \\ 2\eta_M \mathbf{d}(\mathbf{w}_M) \\ 2\eta_H \mathbf{d}(\mathbf{w}_H) \end{bmatrix}. \end{aligned} \tag{17}$$

Each element δ_j of $\mathbf{\Delta}$ is bounded by $\pm 2^{-(\beta_p+1)}$, that is, $\|\mathbf{\Delta}\|_{\max} < 2^{-(\beta_p+1)}$. With the perturbation $\mathbf{\Delta}$, $\lambda_i(\overline{\mathbf{A}}(\mathbf{w}))$ is moved to $\lambda_i(\overline{\mathbf{A}}(\mathbf{w} + \mathbf{r}(\mathbf{w},\alpha) \circ \mathbf{\Delta}))$. If an eigenvalue of $\overline{\mathbf{A}}(\mathbf{w} + \mathbf{r}(\mathbf{w},\alpha) \circ \mathbf{\Delta})$ is outside the open unit disk, the closed-loop system, designed to be stable, becomes unstable with the finite-precision implemented $\mathbf{w}$. Therefore, it is critical to know when the FWL error will cause closed-loop instability. This means that we would like to know the largest open "hypercube" in the perturbation space within which the closed-loop system remains stable. Based on this consideration, a precision measure for realization $\mathbf{w}$ of format α can be defined as

$$\mu_0(\mathbf{w},\alpha) \triangleq \inf\left\{\|\mathbf{\Delta}\|_{\max} : \overline{\mathbf{A}}(\mathbf{w} + \mathbf{r}(\mathbf{w},\alpha) \circ \mathbf{\Delta}) \text{ is unstable}\right\}. \tag{18}$$

From the previous definition, the following proposition is obvious.

Proposition 2: $\overline{\mathbf{A}}(\mathbf{w} + \mathbf{r}(\mathbf{w},\alpha) \circ \mathbf{\Delta})$ is stable if $\|\mathbf{\Delta}\|_{\max} < \mu_0(\mathbf{w},\alpha)$.

Thus, under the condition that the dynamic range is sufficient, that is, $\beta_r \geq \beta_r^{\min}$, the perturbation $\|\mathbf{\Delta}\|_{\max}$ and therefore the precision bit length β_p determines whether the closed-loop remains stable. Let $\beta_p^{\min}$ be the smallest precision bit length such that $\forall \beta_p \geq \beta_p^{\min}$, the closed-loop system is stable with $\mathbf{w}$ implemented by β_p precision bits. The precision measure $\mu_0(\mathbf{w},\alpha)$ provides an estimate of $\beta_p^{\min}$ as $\hat{\beta}_{p0}^{\min}(\mathbf{w},\alpha) \triangleq -\lfloor \log_2 \mu_0(\mathbf{w},\alpha) \rfloor - 1$. It can be seen that $\hat{\beta}_{p0}^{\min} \geq \beta_p^{\min}$.

Define the minimum total bit length required in the implementation of $\mathbf{w}$ as $\beta^{\min} \triangleq \beta_r^{\min} + \beta_p^{\min} + 1$. Clearly, $\mathbf{w}$ implemented with a bit length $\beta \geq \beta^{\min}$ can guarantee a sufficient dynamic range and closed-loop stability. Combining the measures $\gamma(\mathbf{w},\alpha)$ and $\mu_0(\mathbf{w},\alpha)$ results in the following true FWL closed-loop stability measure for the given realization $\mathbf{w}$ with format α

$$\rho_0(\mathbf{w},\alpha) \triangleq \frac{\mu_0(\mathbf{w},\alpha)}{\gamma(\mathbf{w},\alpha)}. \tag{19}$$

An estimate of $\beta^{\min}$ is given by $\rho_0(\mathbf{w},\alpha)$ as $\hat{\beta}_0^{\min}(\mathbf{w},\alpha) \triangleq -\lfloor \log_2 \rho_0(\mathbf{w},\alpha) \rfloor + 1$. It is clear that $\hat{\beta}_0^{\min} \geq \beta^{\min}$. The following proposition summarizes the usefulness of $\rho_0(\mathbf{w},\alpha)$ as a measure for the FWL characteristics of $\mathbf{w}$ in representation format α.

Proposition 3: The controller realization $\mathbf{w}$ implemented in format α with a bit length β can guarantee a sufficient dynamic range and closed-loop stability, if $2^{-\beta+1} \leq \rho_0(\mathbf{w},\alpha)$.

Computing the value of $\mu_0(\mathbf{w},\alpha)$, however, is an unsolved open problem. Thus, the true FWL closed-loop stability measure $\rho_0(\mathbf{w},\alpha)$ has limited practical significance. In the next section, an alternative measure is developed which not only can quantify FWL characteristics of $\mathbf{w}$ in format α but also is computationally tractable.

IV. A Tractable FWL Closed-Loop Stability Measure and its Optimization

First, $\forall i \in \{1,\ldots,m+n\}$

$$\left|\lambda_i\left(\overline{\mathbf{A}}\left(\mathbf{w}+\mathbf{r}(\mathbf{w},\alpha)\circ\boldsymbol{\Delta}\right)\right)\right| = \left|\lambda_i\left(\overline{\mathbf{A}}(\mathbf{w})\right)\right| + \int_{\mathcal{G}} \frac{\partial|\lambda_i|}{\partial\boldsymbol{\Delta}}\,d\boldsymbol{\Delta} \tag{20}$$

where $\mathcal{G}$ is the oriented curve from $\mathbf{0}$ to $\boldsymbol{\Delta}$. For the derivative $(\partial|\lambda_i|)/(\partial\boldsymbol{\Delta}) = [(\partial|\lambda_i|)/(\partial\delta_j)]$, define

$$\left\|\frac{\partial|\lambda_i|}{\partial\boldsymbol{\Delta}}\right\|_1 \triangleq \sum_{j=1}^{N}\left|\frac{\partial|\lambda_i|}{\partial\delta_j}\right|. \tag{21}$$

Further define the precision measure for realization $\mathbf{w}$ in format α

$$\mu_i(\mathbf{w},\alpha) \triangleq \min_{i\in\{1,\cdots,m+n\}} \frac{1-\left|\lambda_i\left(\overline{\mathbf{A}}(\mathbf{w})\right)\right|}{\left\|\frac{\partial|\lambda_i|}{\partial\boldsymbol{\Delta}}\Big|_{\boldsymbol{\Delta}=0}\right\|_1}. \tag{22}$$

Obviously, if $\|\boldsymbol{\Delta}\|_{\max} < \mu_1(\mathbf{w},\alpha)$ and

$$\left|\lambda_i\left(\overline{\mathbf{A}}\left(\mathbf{w}+\mathbf{r}(\mathbf{w},\alpha)\circ\boldsymbol{\Delta}\right)\right)\right| - \left|\lambda_i\left(\overline{\mathbf{A}}(\mathbf{w})\right)\right| \le \|\boldsymbol{\Delta}\|_{\max}\left\|\frac{\partial|\lambda_i|}{\partial\boldsymbol{\Delta}}\Big|_{\boldsymbol{\Delta}=0}\right\|_1 \tag{23}$$

then $|\lambda_i(\overline{\mathbf{A}}(\mathbf{w}+\mathbf{r}(\mathbf{w},\alpha)\circ\boldsymbol{\Delta}))| < 1$ which means that the closed-loop remains stable under the FWL error $\boldsymbol{\Delta}$. As discussed in [5] and [6], the condition (23) is satisfied, provided that $\mu_0(\mathbf{w},\alpha)$ is small enough. The assumption of small $\mu_0(\mathbf{w},\alpha)$ is generally valid, as it does not make much sense to study the FWL effects on the closed-loop stability for those situations where the closed-loop systems have a very large stability robustness. Hence, (23) is not restrictive. Thus, with a sufficient dynamic range, the closed-loop can tolerate those FWL perturbations $\boldsymbol{\Delta}$ whose norms $\|\boldsymbol{\Delta}\|_{\max}$ are less than $\mu_1(\mathbf{w},\alpha)$. Similar to $\mu_0(\mathbf{w},\alpha)$, from the precision measure $\mu_1(\mathbf{w},\alpha)$, an estimate of $\beta_p^{\min}$ is given as $\hat{\beta}_{p1}^{\min}(\mathbf{w},\alpha) \triangleq -\lfloor\log_2\mu_1(\mathbf{w},\alpha)\rfloor - 1$.

Comment: In (20), $\mathcal{G}$ should be chosen to avoid those points where derivative $(\partial|\lambda_i|)/(\partial\boldsymbol{\Delta})$ do not exist, and the derivative $(\partial|\lambda_i|)/(\partial\boldsymbol{\Delta})|_{\boldsymbol{\Delta}=0}$ must exist. From the results of [5] and [6], $(\partial|\lambda_i|)/(\partial\boldsymbol{\Delta})|_{\boldsymbol{\Delta}=0}$ exist if $\overline{\mathbf{A}}(\mathbf{w})$ has $m+n$ distinct nonzero eigenvalues. If $\overline{\mathbf{A}}(\mathbf{w})$ has multiple repeating closed-loop eigenvalues, some of $(\partial|\lambda_i|)/(\partial\boldsymbol{\Delta})|_{\boldsymbol{\Delta}=0}$ may not exist, and in this case $\mu_1(\mathbf{w},\alpha)$ is not defined. However, in practical control system designs, it is very rare that $\overline{\mathbf{A}}(\mathbf{w})$ has multiple repeating eigenvalues. As for the case of $\lambda_i = 0$, since the zero eigenvalue has the largest stability margin $1-|\lambda_i|$, it is harder to move across the unit circle under the FWL effects, compared with the other nonzero eigenvalues. Hence, for those $\overline{\mathbf{A}}(\mathbf{w})$ having zero eigenvalue, $\mu_1(\mathbf{w},\alpha)$ may be modified such that it only minimizes $(1-|\lambda_i(\overline{\mathbf{A}}(\mathbf{w}))|)/(\|\partial(|\lambda_i|)/(\partial\boldsymbol{\Delta})|_{\boldsymbol{\Delta}=0}\|_1)$ for those nonzero eigenvalues. Alternatively, the more conservative measures of [4] and [5] could be used for cases where there are zero eigenvalues.

Obviously, $\mu_1(\mathbf{w},\alpha)$ is an approximation of $\mu_0(\mathbf{w},\alpha)$. However, unlike the measure $\mu_0(\mathbf{w},\alpha)$, the value of $\mu_1(\mathbf{w},\alpha)$ can be computed explicitly. It is easy to see that $(\partial|\lambda_i|)/(\partial\boldsymbol{\Delta})|_{\boldsymbol{\Delta}=0} = \mathbf{r}(\mathbf{w},\alpha)\circ(\partial|\lambda_i|)/(\partial\mathbf{w})$ and from the results of [6], it can be shown that

$$\frac{\partial\left|\lambda_i\left(\overline{\mathbf{A}}(\mathbf{w})\right)\right|}{\partial\mathbf{F}} = [\mathbf{0}\ \mathbf{I}]\mathbf{L}_i(\mathbf{w})\begin{bmatrix}\mathbf{0}\\ \mathbf{I}\end{bmatrix}$$

$$\frac{\partial\left|\lambda_i\left(\overline{\mathbf{A}}(\mathbf{w})\right)\right|}{\partial\mathbf{G}} = [\mathbf{0}\ \mathbf{I}]\mathbf{L}_i(\mathbf{w})\begin{bmatrix}\mathbf{C}^T\\ \mathbf{0}\end{bmatrix}$$

$$\frac{\partial\left|\lambda_i\left(\overline{\mathbf{A}}(\mathbf{w})\right)\right|}{\partial\mathbf{J}} = [\mathbf{B}^T\ \mathbf{H}^T]\mathbf{L}_i(\mathbf{w})\begin{bmatrix}\mathbf{0}\\ \mathbf{I}\end{bmatrix}$$

$$\frac{\partial\left|\lambda_i\left(\overline{\mathbf{A}}(\mathbf{w})\right)\right|}{\partial\mathbf{M}} = [\mathbf{B}^T\ \mathbf{H}^T]\mathbf{L}_i(\mathbf{w})\begin{bmatrix}\mathbf{C}^T\\ \mathbf{0}\end{bmatrix}$$

$$\frac{\partial\left|\lambda_i\left(\overline{\mathbf{A}}(\mathbf{w})\right)\right|}{\partial\mathbf{H}} = [\mathbf{0}\ \mathbf{I}]\mathbf{L}_i(\mathbf{w})\begin{bmatrix}\mathbf{C}^T\mathbf{M}^T\\ \mathbf{J}^T\end{bmatrix} \tag{24}$$

with

$$\mathbf{L}_i(\mathbf{w}) \triangleq \frac{\mathrm{Re}\left[\lambda_i^*\left(\overline{\mathbf{A}}(\mathbf{w})\right)\mathbf{y}_i^*\left(\overline{\mathbf{A}}(\mathbf{w})\right)\mathbf{p}_i^T\left(\overline{\mathbf{A}}(\mathbf{w})\right)\right]}{\left|\lambda_i\left(\overline{\mathbf{A}}(\mathbf{w})\right)\right|} \tag{25}$$

where $\mathbf{p}_i(\overline{\mathbf{A}}(\mathbf{w}))$ and $\mathbf{y}_i(\overline{\mathbf{A}}(\mathbf{w}))$ are the right and reciprocal left eigenvectors related to $\lambda_i(\overline{\mathbf{A}}(\mathbf{w}))$, respectively, * denotes the conjugate operation and $\mathrm{Re}[\cdot]$ the real part. Replacing $\mu_0(\mathbf{w},\alpha)$ with $\mu_1(\mathbf{w},\alpha)$ in (19) leads to a computationally tractable FWL closed-loop stability measure

$$\rho_1(\mathbf{w},\alpha) \triangleq \frac{\mu_1(\mathbf{w},\alpha)}{\gamma(\mathbf{w},\alpha)}. \tag{26}$$

From $\rho_1(\mathbf{w},\alpha)$, an estimate of $\beta^{\min}$ is given as $\hat{\beta}_1^{\min}(\mathbf{w},\alpha) \triangleq -\lfloor\log_2\rho_1(\mathbf{w},\alpha)\rfloor + 1$. Compared with the existing FWL measures [1]–[8], $\rho_1(\mathbf{w},\alpha)$ has at least two advantages. First, $\rho_1(\mathbf{w},\alpha)$ can be used in different representation formats while the existing measures are only valid for a particular format. For example, the measures presented in [3]–[6] are fixed-point measures and the measure in [7] is a floating-point one. The measure $\rho_1(\mathbf{w},\alpha)$ offers a unified framework to compare the FWL characteristics of a realization $\mathbf{w}$ in different formats. Second and more critically, unlike the existing measures which are precision measures only and imply an unlimited dynamic range, $\rho_1(\mathbf{w},\alpha)$ is made up of a dynamic range measure and a precision measure and is therefore a true FWL measure capable of handling closed-loop stability as well as the underflow and overflow aspects.

In a given format α, different realizations $\mathbf{w}$ yield different values of $\rho_1(\mathbf{w},\alpha)$. It is of practical importance to find an "optimal" realization $\mathbf{w}_{\mathrm{opt}}(\alpha)$ that maximizes $\rho_1(\mathbf{w},\alpha)$ for the format α. The controller implemented with this optimal realization $\mathbf{w}_{\mathrm{opt}}(\alpha)$ in format α needs a minimum bit length and has a maximum tolerance to the FWL error. This optimal realization problem is formally defined as

$$v(\alpha) \triangleq \max_{\mathbf{w}\in\mathcal{S}_C}\rho_1(\mathbf{w},\alpha). \tag{27}$$

Considering that $\mathbf{w}$ is a function of $\mathbf{T}$, $\mathbf{r}(\mathbf{w},\alpha)$ and $\gamma(\mathbf{w},\alpha)$ depend on $\mathbf{T}$ and α, we can define the following optimization criterion in format α:

$$\xi(\mathbf{T},\alpha) \triangleq \min_{i\in\{1,\cdots,m+n\}} \frac{1-\left|\lambda_i\left(\overline{\mathbf{A}}(\mathbf{w}_0)\right)\right|}{\left\|\mathbf{r}(\mathbf{w},\alpha)\circ\frac{\partial|\lambda_i|}{\partial\mathbf{w}}\right\|_1\gamma(\mathbf{w},\alpha)} = \rho_1(\mathbf{w},\alpha). \tag{28}$$

The optimal realization problem (27) can then be posed as the following optimization problem:

$$v(\alpha) = \max_{\substack{\mathbf{T}\in\mathcal{R}^{m\times m}\\ \det(\mathbf{T})\neq 0}}\xi(\mathbf{T},\alpha). \tag{29}$$

As the optimization problem (29) is highly nonlinear, global optimization algorithms, such as the genetic algorithm [9] and adaptive simulated annealing [10], can be adopted to provide a (sub)optimal similarity transformation $\mathbf{T}_{\mathrm{opt}}(\alpha)$. Global optimization methods are however computationally demanding. Local optimization algorithms, such

as Rosenbrock and Simplex algorithms [11], are computationally simpler but run more risks of only attaining a local solution. Our experience with the optimization problem (29) suggests that, unlike optimizing the precision measure $\mu_1(\mathbf{w},1)$ alone [6], the dynamic range measure $\gamma(\mathbf{w},\alpha)$ in the criterion $\rho_1(\mathbf{w},\alpha)$ helps to bound the solution set and the cost function $\xi(\mathbf{T},\alpha)$ appears to behave better. It also help to choose a "good" initial controller realization, such as the open-loop balanced realization [12] or Li's closed-loop suboptimal realization [4], as the initial guess for the optimization routine.

With $\mathbf{T}_{\rm opt}(\alpha)$, the corresponding optimal realization $\mathbf{w}_{\rm opt}(\alpha)$ in format α can readily be computed. By setting $\alpha = 1, 2, 3$, respectively, in the optimization problem (29), we can attain an optimal fixed-point realization $\mathbf{w}_{\rm opt}(1)$, an optimal floating-point realization $\mathbf{w}_{\rm opt}(2)$ and an optimal block-floating-point realization $\mathbf{w}_{\rm opt}(3)$ for a digital controller. It is worth reiterating that the optimization problem (29) yields a true optimal controller realization, as the solution $\mathbf{T}_{\rm opt}(\alpha)$ minimizes the required β_p as well as β_r and, therefore, minimizes the required total bit length β. This should be compared with the existing "optimal" realization problems [1]–[8], which only try to minimize the required precision bit β_p and, as a consequence, do not necessarily minimize the required total bit length β.

It is interesting to compare our approach with eigenstructure orthogonalization, which is also based on eigenvalue sensitivities [1]. For a complex-valued matrix $\mathbf{U}$, let $\|\mathbf{U}\|_2$ represent its largest singular value. The following lemma summarizes three properties of $\|\cdot\|_2$.

Lemma 1: $\|\mathbf{U}\|_2 \geq \|\mathrm{Re}[\mathbf{U}]\|_2$; $\left\|\begin{bmatrix}\mathbf{U}_1\\ \mathbf{U}_2\end{bmatrix}\right\|_2 \geq \|\mathbf{U}_1\|_2$; $\|[\mathbf{U}_1 \quad \mathbf{U}_2]\|_2 \geq \|\mathbf{U}_1\|_2$.

For an illustrative purpose, we consider the case of $\alpha = 1$ (fixed-point format) with $\mathbf{F}$, $\mathbf{G}$, $\mathbf{J}$ and $\mathbf{M}$ being nonzero matrices and $\mathbf{H} = \mathbf{0}$ in (7). Denote the controller realization

$$\mathbf{X} \triangleq \begin{bmatrix}\mathbf{M} & \mathbf{J}\\ \mathbf{G} & \mathbf{F}\end{bmatrix}. \tag{30}$$

In this case, $\mathbf{X}$ is perturbed to $\mathbf{X} + \mathbf{\Delta}_X$ due to the FWL effects, and

$$\begin{aligned}\left.\frac{\partial|\lambda_i|}{\partial \mathbf{\Delta}_X}\right|_{\mathbf{\Delta}_X=\mathbf{0}} &= \frac{\partial|\lambda_i|}{\partial \mathbf{X}} = \begin{bmatrix}\frac{\partial|\lambda_i|}{\partial \mathbf{M}} & \frac{\partial|\lambda_i|}{\partial \mathbf{J}}\\ \frac{\partial|\lambda_i|}{\partial \mathbf{G}} & \frac{\partial|\lambda_i|}{\partial \mathbf{F}}\end{bmatrix}\\ &= \begin{bmatrix}\mathbf{B}^T & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}\mathbf{L}_i(\mathbf{w})\begin{bmatrix}\mathbf{C}^T & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}.\end{aligned} \tag{31}$$

Applying Lemma 1 to (25) brings about

$$\begin{aligned}\|\mathbf{L}_i(\mathbf{w})\|_2 &= \frac{\left\|\mathrm{Re}\left[\lambda_i^*\left(\overline{\mathbf{A}}(\mathbf{w})\right)\mathbf{y}_i^*\left(\overline{\mathbf{A}}(\mathbf{w})\right)\mathbf{p}_i^T\left(\overline{\mathbf{A}}(\mathbf{w})\right)\right]\right\|_2}{\left|\lambda_i\left(\overline{\mathbf{A}}(\mathbf{w})\right)\right|}\\ &\leq \frac{\left\|\lambda_i^*\left(\overline{\mathbf{A}}(\mathbf{w})\right)\mathbf{y}_i^*\left(\overline{\mathbf{A}}(\mathbf{w})\right)\mathbf{p}_i^T\left(\overline{\mathbf{A}}(\mathbf{w})\right)\right\|_2}{\left|\lambda_i\left(\overline{\mathbf{A}}(\mathbf{w})\right)\right|}\\ &\leq \left\|\mathbf{y}_i^*\left(\overline{\mathbf{A}}(\mathbf{w})\right)\right\|_2\left\|\mathbf{p}_i^T\left(\overline{\mathbf{A}}(\mathbf{w})\right)\right\|_2.\end{aligned} \tag{32}$$

Then

$$\begin{aligned}\left\|\left.\frac{\partial|\lambda_i|}{\partial \mathbf{\Delta}_X}\right|_{\mathbf{\Delta}_X=\mathbf{0}}\right\|_2 &\leq \left\|\begin{bmatrix}\mathbf{B}^T & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}\right\|_2\|\mathbf{L}_i(\mathbf{w})\|_2\left\|\begin{bmatrix}\mathbf{C}^T & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}\right\|_2\\ &\leq \varphi\left\|\mathbf{y}_i^*\left(\overline{\mathbf{A}}(\mathbf{w})\right)\right\|_2\left\|\mathbf{p}_i^T\left(\overline{\mathbf{A}}(\mathbf{w})\right)\right\|_2\end{aligned} \tag{33}$$

where

$$\varphi \triangleq \left\|\begin{bmatrix}\mathbf{B}^T & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}\right\|_2\left\|\begin{bmatrix}\mathbf{C}^T & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}\right\|_2. \tag{34}$$

Applying Lemma 1 to (33) for the $m+n$ eigenvalues results in

$$\max_{i\in\{1,\cdots,m+n\}}\left\|\left.\frac{\partial|\lambda_i|}{\partial \mathbf{\Delta}_X}\right|_{\mathbf{\Delta}_X=\mathbf{0}}\right\|_2 \leq \varphi\|\mathbf{Y}\|_2\|\mathbf{P}\|_2 \tag{35}$$

where

$$\mathbf{Y} \triangleq [\mathbf{y}_1^*(\overline{\mathbf{A}}(\mathbf{w}))\ldots\mathbf{y}_{m+n}^*(\overline{\mathbf{A}}(\mathbf{w}))]$$

and

$$\mathbf{P} \triangleq [\mathbf{p}_1(\overline{\mathbf{A}}(\mathbf{w}))\ldots\mathbf{p}_{m+n}(\overline{\mathbf{A}}(\mathbf{w}))]^T.$$

Noting the relationship $\mathbf{Y} = \mathbf{P}^{-1}$ between right eigenvectors and left eigenvectors, we can see that

$$\max_{i\in\{1,\cdots,m+n\}}\left\|\left.\frac{\partial|\lambda_i|}{\partial \mathbf{\Delta}_X}\right|_{\mathbf{\Delta}_X=\mathbf{0}}\right\|_2 \leq \varphi\|\mathbf{P}^{-1}\|_2\|\mathbf{P}\|_2 \tag{36}$$

which gives an upper bound of the sensitivities of the eigenvalues. Based on (36), making the eigenvalues insensitive needs to find those eigenvectors $\mathbf{P}$ which minimize $\kappa(\mathbf{P}) \triangleq \|\mathbf{P}^{-1}\|_2\|\mathbf{P}\|_2$. The results of [13] show that if and only if $\mathbf{P}$ is a normal matrix, $\kappa(\mathbf{P})$ takes the minimal value. If this happens, $\mathbf{P}$ and $\mathbf{Y}$ can be scaled to give an orthonormal basis of $\mathcal{C}^n$, and $\kappa(\mathbf{P}) = 1$. This is the idea of eigenstructure orthogonalization for finding the realizations which have closed-loop eigenvalues of low sensitivities.

A comparison of our approach with eigenstructure orthogonalization can now be made. Firstly, our approach directly adopts the eigenvalue sensitivities while eigenstructure orthogonalization adopts the bound $\kappa(\mathbf{P})$ of the eigenvalue sensitivities. This implies that eigenstructure orthogonalization is conservative in comparison with our approach. Secondly, our approach considers both the stability margins and the eigenvalue sensitivities in (22) and is, therefore, able to evaluate the FWL stability of a system accurately while eigenstructure orthogonalization only considers a bound of the eigenvalue sensitivities and cannot provide any estimate of the required bit length. Finally, it should be pointed out that for most practical systems, owing to the limited degrees of freedom, there does not exist any feasible controller realization achieving orthogonal closed-loop eigenstructure. Thus, for the purpose of minimizing $\kappa(\mathbf{P})$, the eigenstructure assignment techniques (see for example [14]) are employed instead to choose eigenvectors which are as mutually orthogonal as possible. The resulting realizations are obviously more conservative.

V. Design Examples and Result Comparison

Example 1: This example was taken from [6]. The discrete-time plant was given by (37), shown at the bottom of the next page. The initial realization of the digital controller was given by

$$\begin{aligned}\mathbf{F}_0 &= \begin{bmatrix}0 & 1.0000e+0\\ -9.3303e-1 & 1.9319e+0\end{bmatrix}\\ \mathbf{G}_0 &= \begin{bmatrix}4.1814e-2 & 2.7132e+2\\ 3.9090e-2 & 1.0167e+3\end{bmatrix}\\ \mathbf{J}_0 &= [3.0000e-4 \quad 5.0000e-4]\\ \mathbf{M}_0 &= [0 \quad 6.1250e-1]\\ \mathbf{H}_0 &= \begin{bmatrix}7.8047e+1\\ 7.3849e+1\end{bmatrix}.\end{aligned}$$

Based on the proposed unified FWL closed-loop stability measure, the optimization problem (29) was formed. Using the MATLAB routine *fminsearch.m*, which is a local optimization routine, this optimization problem was solved for $\alpha = 1, 2, 3$, respectively, to obtain the optimal similarity transformation in fixed-point format $\mathbf{T}_{\rm opt}(1)$, the optimal similarity transformation in floating-point format $\mathbf{T}_{\rm opt}(2)$ and the optimal similarity transformation in block-floating-point format $\mathbf{T}_{\rm opt}(3)$.

TABLE I
VARIOUS MEASURES AND ESTIMATED BIT LENGTHS FOR THE FOUR REALIZATIONS IN THREE DIFFERENT FORMATS OF EXAMPLE 1

	$\mathbf{w}_0$	$\mathbf{w}_{\rm opt}(1)$	$\mathbf{w}_{\rm opt}(2)$	$\mathbf{w}_{\rm opt}(3)$
$\rho_1(\mathbf{w},1)$	$\mathbf{2.5150e-9}$	$\mathbf{1.1386e-7}$	$2.7728e-8$	$1.0861e-7$
$\hat{\beta}_1^{min}(\mathbf{w},1)$	**30**	**25**	27	25
$\mu_1(\mathbf{w},1)$	$\mathbf{2.5569e-6}$	$\mathbf{5.0795e-7}$	$2.5937e-5$	$1.7450e-7$
$\hat{\beta}_{p1}^{min}(\mathbf{w},1)$	**18**	**20**	15	22
$\gamma(\mathbf{w},1)$	$\mathbf{1.0167e+3}$	$\mathbf{4.4612e+0}$	$9.3543e+2$	$1.6066e+0$
$\hat{\beta}_r^{min}(\mathbf{w},1)$	**10**	**3**	10	1
$\rho_1(\mathbf{w},2)$	$\mathbf{1.3134e-7}$	$1.9204e-5$	$\mathbf{1.9593e-5}$	$3.3365e-7$
$\hat{\beta}_1^{min}(\mathbf{w},2)$	**24**	17	**17**	23
$\mu_1(\mathbf{w},2)$	$\mathbf{3.1118e-6}$	$4.3127e-4$	$\mathbf{4.3127e-4}$	$5.4490e-6$
$\hat{\beta}_{p1}^{min}(\mathbf{w},2)$	**18**	11	**11**	17
$\gamma(\mathbf{w},2)$	$\mathbf{2.3692e+1}$	$2.2458e+1$	$\mathbf{2.2012e+1}$	$1.6332e+1$
$\hat{\beta}_r^{min}(\mathbf{w},2)$	**5**	5	**5**	5
$\rho_1(\mathbf{w},3)$	$\mathbf{9.2976e-10}$	$5.3779e-9$	$2.8185e-9$	$\mathbf{1.3362e-8}$
$\hat{\beta}_1^{min}(\mathbf{w},3)$	**32**	29	30	**28**
$\mu_1(\mathbf{w},3)$	$\mathbf{2.1343e-8}$	$5.7385e-8$	$5.7266e-8$	$\mathbf{5.4549e-8}$
$\hat{\beta}_{p1}^{min}(\mathbf{w},3)$	**25**	24	24	**24**
$\gamma(\mathbf{w},3)$	$\mathbf{2.2955e+1}$	$1.0671e+1$	$2.0318e+1$	$\mathbf{4.0823e+0}$
$\hat{\beta}_r^{min}(\mathbf{w},3)$	**5**	4	5	**3**

TABLE II
VARIOUS MEASURES AND ESTIMATED BIT LENGTHS FOR THE FOUR REALIZATIONS IN THREE DIFFERENT FORMATS OF EXAMPLE 2

	$\mathbf{w}_0$	$\mathbf{w}_{\rm opt}(1)$	$\mathbf{w}_{\rm opt}(2)$	$\mathbf{w}_{\rm opt}(3)$
$\rho_1(\mathbf{w},1)$	$\mathbf{1.2312e-10}$	$\mathbf{1.2003e-6}$	$1.0580e-7$	$1.1321e-6$
$\hat{\beta}_1^{min}(\mathbf{w},1)$	**34**	**21**	25	21
$\mu_1(\mathbf{w},1)$	$\mathbf{3.3474e-8}$	$\mathbf{2.3082e-4}$	$9.6673e-5$	$2.2287e-4$
$\hat{\beta}_{p1}^{min}(\mathbf{w},1)$	**24**	**12**	13	12
$\gamma(\mathbf{w},1)$	$\mathbf{2.7188e+2}$	$\mathbf{1.9231e+2}$	$9.1370e+2$	$1.9687e+2$
$\hat{\beta}_r^{min}(\mathbf{w},1)$	**9**	**8**	10	8
$\rho_1(\mathbf{w},2)$	$\mathbf{2.9062e-11}$	$7.6826e-6$	$\mathbf{9.5931e-6}$	$8.5778e-6$
$\hat{\beta}_1^{min}(\mathbf{w},2)$	**37**	18	**18**	18
$\mu_1(\mathbf{w},2)$	$\mathbf{2.2389e-10}$	$9.5628e-5$	$\mathbf{1.5229e-4}$	$1.1822e-4$
$\hat{\beta}_{p1}^{min}(\mathbf{w},2)$	**32**	13	**12**	13
$\gamma(\mathbf{w},2)$	$\mathbf{7.7038e+0}$	$1.2447e+1$	$\mathbf{1.5875e+1}$	$1.3782e+1$
$\hat{\beta}_r^{min}(\mathbf{w},2)$	**3**	4	**4**	4
$\rho_1(\mathbf{w},3)$	$\mathbf{1.4347e-11}$	$3.2975e-6$	$3.6938e-7$	$\mathbf{3.5012e-6}$
$\hat{\beta}_1^{min}(\mathbf{w},3)$	**38**	20	23	**20**
$\mu_1(\mathbf{w},3)$	$\mathbf{6.5127e-11}$	$2.7666e-5$	$2.9985e-6$	$\mathbf{3.0083e-5}$
$\hat{\beta}_{p1}^{min}(\mathbf{w},3)$	**33**	15	18	**15**
$\gamma(\mathbf{w},3)$	$\mathbf{4.5395e+0}$	$8.3902e+0$	$8.1176e+0$	$\mathbf{8.5923e+0}$
$\hat{\beta}_r^{min}(\mathbf{w},3)$	**3**	4	4	**4**

These in turn provided the three corresponding optimal controller realizations $\mathbf{w}_{\rm opt}(1)$, $\mathbf{w}_{\rm opt}(2)$, and $\mathbf{w}_{\rm opt}(3)$.

Example 2: In this example, the discrete-time plant taken from [1] was given by

$$\mathbf{A}=\begin{bmatrix} 3.7156e+0 & -5.4143e+0 & 3.6525e+0 & -9.6420e-1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

$$\mathbf{B}=[1 \quad 0 \quad 0 \quad 0]^T$$

$$\mathbf{C}=[1.1160e-6 \quad 4.3000e-8 \quad 1.0880e-6 \quad 1.4000e-8].$$

The initial realization of the digital controller, which was a modification of the initial output-feedback controller in [1] by a similarity transformation, was given by (38), shown at the bottom of the page. Using the same method for Example 1, the three optimal controller realizations $\mathbf{w}_{\rm opt}(\alpha)$ were obtained for $\alpha = 1, 2, 3$.

Tables I and II list, for Examples 1 and 2, respectively, the values of the measures ρ_1, μ_1, and γ for the three different representation schemes together with the corresponding estimated bit lengths for the initial realization $\mathbf{w}_0$, the optimal fixed-point realization $\mathbf{w}_{\rm opt}(1)$, the optimal floating-point realization $\mathbf{w}_{\rm opt}(2)$ and the optimal block-floating-point realization $\mathbf{w}_{\rm opt}(3)$. In these two tables, the various estimated bit lengths were computed from their respective measure values. Some observations can readily be made from the results in Tables I and and II.

As far as the robustness of FWL closed-loop stability is concerned, given an arbitrary realization, floating-point representation is not necessarily better than fixed-point or block-floating-point one. For example, floating-point is the best format to implement the initial real-

$$\mathbf{A}=\begin{bmatrix} 3.2439e-1 & -4.5451e+0 & -4.0535e+0 & -2.7003e-3 & 0 \\ 1.4518e-1 & 4.9477e-1 & -4.6945e-1 & -3.1274e-4 & 0 \\ 1.6814e-2 & 1.6491e-1 & 9.6681e-1 & -2.2114e-5 & 0 \\ 1.1889e-3 & 1.8209e-2 & 1.9829e-1 & 1.0000e+0 & 0 \\ 6.1301e-5 & 1.2609e-3 & 1.9930e-2 & 2.0000e-1 & 1.0000e+0 \end{bmatrix}$$

$$\mathbf{B}=[1.4518e-1 \quad 1.6814e-2 \quad 1.1889e-3 \quad 6.1301e-5 \quad 2.4979e-6]^T$$

$$\mathbf{C}=\begin{bmatrix} 0 & 0 & 1.6188e+0 & -1.5750e-1 & -4.3943e+1 \\ 1.0000e+0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{37}$$

$$\mathbf{F}_0=\begin{bmatrix} 2.6963e+2 & -4.2709e+1 & 2.2873e+1 & 2.6184e+2 \\ 2.5561e+2 & -4.0497e+1 & 2.1052e+1 & 2.4806e+2 \\ 5.6096e+1 & -8.5715e+0 & 5.2162e+0 & 5.4920e+1 \\ -2.3907e+2 & 3.7998e+1 & -2.0338e+1 & -2.3203e+2 \end{bmatrix}$$

$$\mathbf{G}_0=\begin{bmatrix} -4.6765e+1 \\ -4.5625e+1 \\ -9.5195e+0 \\ 4.1609e+1 \end{bmatrix}$$

$$\mathbf{J}_0=[-2.5548e+2 \quad -2.7185e+2 \quad -2.7188e+2 \quad 2.7188e+2]$$

$$\mathbf{M}_0=[0]$$

$$\mathbf{H}_0=[0 \quad 0 \quad 0 \quad 0]^T. \tag{38}$$

TABLE III
TRUE MINIMUM REQUIRED BIT LENGTHS FOR THE FOUR REALIZATIONS IN DIFFERENT FORMATS OF EXAMPLE 1

Realization	Format	β^{min}	β_p^{min}	β_r^{min}
$\mathbf{w}_0$	fixed-point	23	12	10
$\mathbf{w}_{\rm opt}(1)$	fixed-point	22	18	3
$\mathbf{w}_0$	floating-point	16	10	5
$\mathbf{w}_{\rm opt}(2)$	floating-point	12	6	5
$\mathbf{w}_0$	block-floating-point	28	22	5
$\mathbf{w}_{\rm opt}(3)$	block-floating-point	23	20	2

TABLE IV
TRUE MINIMUM REQUIRED BIT LENGTHS FOR THE FOUR REALIZATIONS IN DIFFERENT FORMATS OF EXAMPLE 2

Realization	Format	β^{min}	β_p^{min}	β_r^{min}
$\mathbf{w}_0$	fixed-point	31	21	9
$\mathbf{w}_{\rm opt}(1)$	fixed-point	19	10	8
$\mathbf{w}_0$	floating-point	33	29	3
$\mathbf{w}_{\rm opt}(2)$	floating-point	13	8	4
$\mathbf{w}_0$	block-floating-point	33	30	2
$\mathbf{w}_{\rm opt}(3)$	block-floating-point	16	12	3

ization $\mathbf{w}_0$ of Example 1 while fixed-point is the best format to implement $\mathbf{w}_0$ of Example 2. In fact, for Example 2, we had deliberately chosen $\mathbf{w}_0$ as the transformation of the initial controller realization in [1] by a similarity transformation matrix to favor a fixed-point implementation. However, as expected, the optimal floating-point realization $\mathbf{w}_{\rm opt}(2)$ implemented in floating-point format is always the best in terms of robustness to FWL errors. Also, the results in Table I show that fixed-point format is better than block-floating-point format to implement $\mathbf{w}_{\rm opt}(\alpha)$ of Example 1 for $1 \leq \alpha \leq 3$, while the results of Table II indicate that the opposite is true for Example 2. This simply confirms the fact that the performance of block-floating-point scheme critically depends on how to divide $\mathbf{w}$ into blocks. With a proper division, block-floating-point scheme should beat fixed-point scheme in terms of robustness to FWL errors. The results also show that the proposed optimization procedure is very effective. This can be seen by comparing the values of the measure for $\mathbf{w}_0$ and $\mathbf{w}_{\rm opt}(\alpha)$ implemented in a same format α.

Table III compares the true minimum required bit lengths $\beta_r^{\rm min}$, $\beta_p^{\rm min}$ and $\beta^{\rm min}$ of the initial realization $\mathbf{w}_0$ implemented in the three different schemes with those of fixed-point implemented $\mathbf{w}_{\rm opt}(1)$, floating-point implemented $\mathbf{w}_{\rm opt}(2)$ and block-floating-point implemented $\mathbf{w}_{\rm opt}(3)$ of Example 1. It can be seen that the floating-point implemented $\mathbf{w}_{\rm opt}(2)$ requires at least 12 bits to ensure closed-loop stability which is much better than minimum 22 bits needed by fixed-point implemented $\mathbf{w}_{\rm opt}(1)$ or minimum 23 bits needed by block-floating-point implemented $\mathbf{w}_{\rm opt}(3)$. Table IV summarizes the minimum required bit lengths $\beta_r^{\rm min}$, $\beta_p^{\rm min}$, and $\beta^{\rm min}$ for fixed-point implemented $\mathbf{w}_{\rm opt}(1)$, floating-point implemented $\mathbf{w}_{\rm opt}(2)$ and block-floating-point implemented $\mathbf{w}_{\rm opt}(3)$ of Example 2 together with those for $\mathbf{w}_0$ in the three formats. It can be seen that the floating-point implemented $\mathbf{w}_{\rm opt}(2)$ needs at least 13 bits to maintain closed-loop stability which is again better than minimum 19 bits needed by fixed-point implemented $\mathbf{w}_{\rm opt}(1)$ or minimum 16 bits needed by block-floating-point implemented $\mathbf{w}_{\rm opt}(3)$.

Notice that any realization $\mathbf{w} \in \mathcal{S}_C$ implemented in infinite precision (unlimited β_r and infinite β_p) will achieve the exact performance of the infinite-precision implemented $\mathbf{w}_0$, which is the designed controller performance. For this reason, the infinite-precision implemented $\mathbf{w}_0$ is referred to as the ideal controller realization $\mathbf{w}_{\rm ideal}$. In Example 1, there are two outputs $\mathbf{y}(k) = [y_1(k)y_2(k)]^T$. Fig. 1 compares the unit impulse response of the first plant output $y_1(k)$ of Example 1 for the ideal controller $\mathbf{w}_{\rm ideal}$ with those of the 15-bit floating-point implemented $\mathbf{w}_0$ (five exponent bits and nine mantissa bits) and the 15-bit floating-point implemented $\mathbf{w}_{\rm opt}(2)$ (five exponent bits and nine mantissa bits). Fig. 2 compares the unit impulse response of the plant output $y(k)$ of Example 2 for $\mathbf{w}_{\rm ideal}$ with those of the 33-bit block-floating-point implemented $\mathbf{w}_0$ (two block exponent bits and 30 block mantissa bits) and the 33-bit block-floating-point implemented $\mathbf{w}_{\rm opt}(3)$ (three block exponent bits and 29 block mantissa bits). These results clearly show that, for a chosen α, the corresponding optimal realization is always much better than the initial realization.

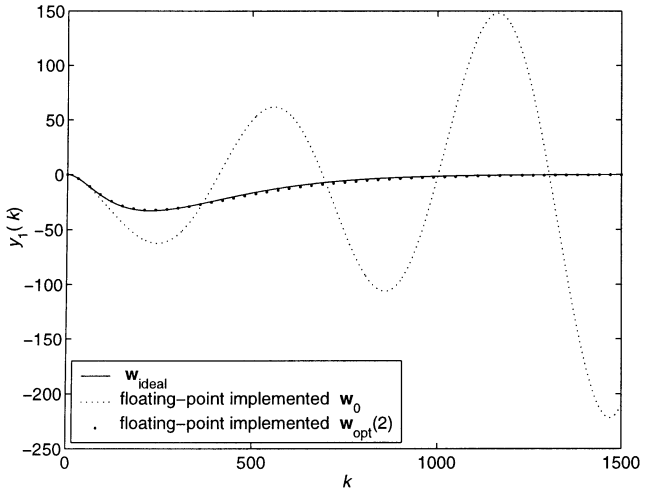

Fig. 1. Unit impulse response of $y_1(k)$ for $\mathbf{w}_{\rm ideal}$, 15-bit floating-point implemented $\mathbf{w}_0$ (five exponent bits and nine mantissa bits), and 15-bit floating-point implemented $\mathbf{w}_{\rm opt}(2)$ (five exponent bits and nine mantissa bits) of Example 1.

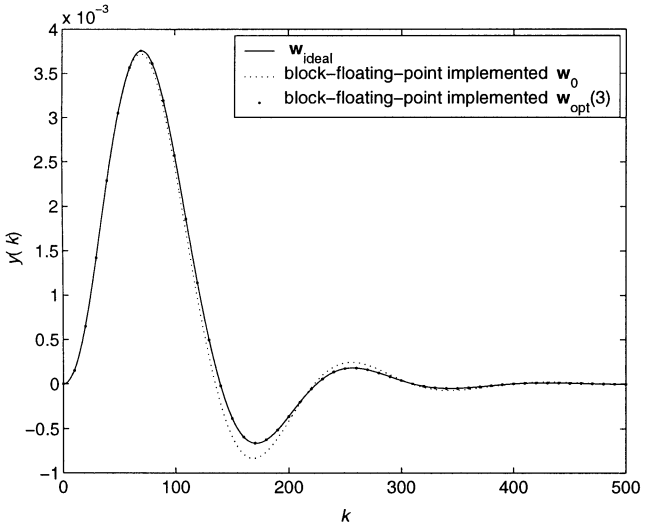

Fig. 2. Unit impulse response of $y(k)$ for $\mathbf{w}_{\rm ideal}$, 33-bit block-floating-point implemented $\mathbf{w}_0$ (two block exponent bits and 30 block mantissa bits), and 33-bit block-floating-point implemented $\mathbf{w}_{\rm opt}(3)$ (three block exponent bits and 29 block mantissa bits) of Example 2.

Fig. 3 compares the unit impulse response of $y_1(k)$ of Example 1 for $\mathbf{w}_{\rm ideal}$ with those of the 22-bit fixed-point implemented $\mathbf{w}_{\rm opt}(1)$ ($\beta_g = 3$ and $\beta_f = 18$), the 22-bit floating-point implemented $\mathbf{w}_{\rm opt}(2)$ ($\beta_e = 5$, and $\beta_w = 16$) and the 22-bit block-floating-point implemented $\mathbf{w}_{\rm opt}(3)$ ($\beta_h = 2$ block and $\beta_u = 19$). Fig. 4 compares the unit impulse response of $y(k)$ for $\mathbf{w}_{\rm ideal}$ with those of the 18-bit fixed-point implemented $\mathbf{w}_{\rm opt}(1)$ ($\beta_g = 8$ and $\beta_f = 9$), the 18-bit floating-point

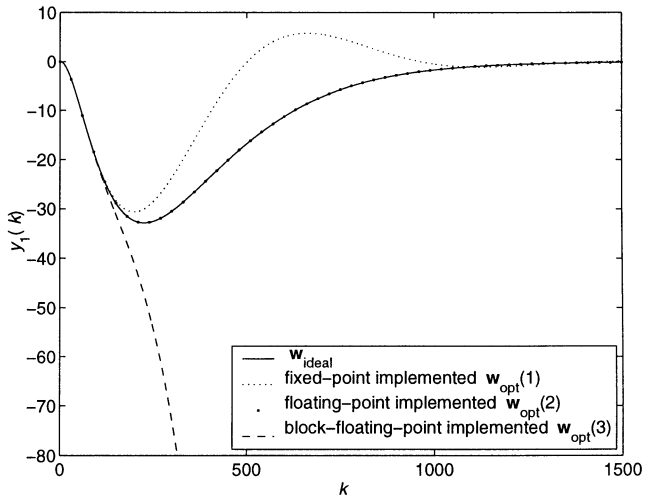

Fig. 3. Unit impulse response of $y_1(k)$ for $\mathbf{w}_{\rm ideal}$, 22-bit fixed-point implemented $\mathbf{w}_{\rm opt}(1)$ (three integer bits and 18 fractional bits), 22-bit floating-point implemented $\mathbf{w}_{\rm opt}(2)$ (five exponent bits and 16 mantissa bits), and 22-bit block-floating-point implemented $\mathbf{w}_{\rm opt}(3)$ (two block exponent bits and 19 block mantissa bits) of Example 1.

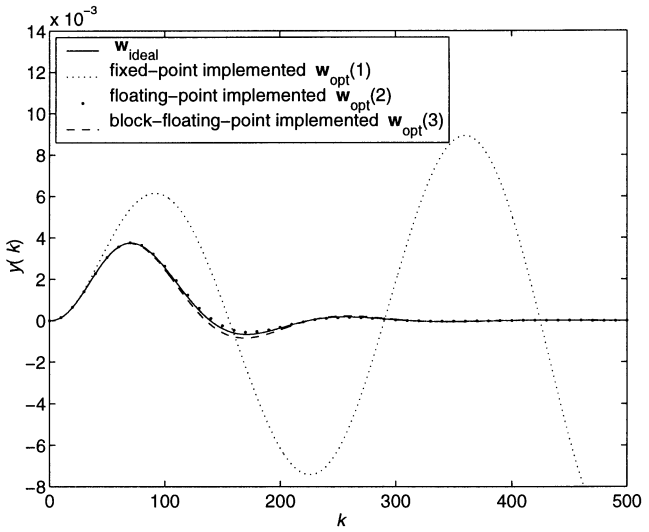

Fig. 4. Unit impulse response of $y(k)$ for $\mathbf{w}_{\rm ideal}$, 18-bit fixed-point implemented $\mathbf{w}_{\rm opt}(1)$ (eight integer bits and nine fractional bits), 18-bit floating-point implemented $\mathbf{w}_{\rm opt}(2)$ (four exponent bits and 13 mantissa bits), and 18-bit block-floating-point implemented $\mathbf{w}_{\rm opt}(3)$ (three block exponent bits and 14 block mantissa bits) of Example 2.

implemented $\mathbf{w}_{\rm opt}(2)$ ($\beta_e = 4$ and $\beta_w = 13$) and the 18-bit block-floating-point implemented $\mathbf{w}_{\rm opt}(3)$ ($\beta_h = 3$ and $\beta_u = 14$) of Example 2. It is obvious from these two figures that the response with floating-point implemented $\mathbf{w}_{\rm opt}(2)$ is the closest to the ideal performance.

VI. Conclusion

We have proposed a design procedure for optimal controller realizations in different representation schemes. The procedure provides designer with useful quantitative information regarding finite precision computational properties, namely robustness to FWL errors and estimated minimum bit length for guaranteeing closed-loop stability. This allows designers to choose an optimal controller realization in an appropriate representation scheme to achieve the best computational efficiency and closed-loop performance.

References

[1] M. Gevers and G. Li, *Parameterizations in Control, Estimation and Filtering Problems: Accuracy Aspects*. London, U.K.: Springer-Verlag, 1993.

[2] R. S. H. Istepanian and J. F. Whidborne, Eds., *Digital Controller Implementation and Fragility: A Modern Perspective*. London, U.K.: Springer-Verlag, 2001.

[3] I. J. Fialho and T. T. Georgiou, "On stability and performance of sampled-data systems subject to wordlength constraint," *IEEE Trans. Automat. Contr.*, vol. 39, pp. 2476–2481, Dec. 1994.

[4] G. Li, "On the structure of digital controllers with finite word length consideration," *IEEE Trans. Automat. Contr.*, vol. 43, pp. 689–693, May 1998.

[5] J. Wu, S. Chen, G. Li, and J. Chu, "Optimal finite-precision state-estimate feedback controller realization of discrete-time systems," *IEEE Trans. Automat. Contr.*, vol. 45, pp. 1550–1554, Aug. 2000.

[6] J. Wu, S. Chen, G. Li, R. S. H. Istepanian, and J. Chu, "An improved closed-loop stability related measure for finite-precision digital controller realizations," *IEEE Trans. Automat. Contr.*, vol. 46, pp. 1162–1166, July 2001.

[7] J. F. Whidborne and D.-W. Gu, "Optimal finite-precision controller and filter realizations using floating-point arithmetic," in *Proc. 15th IFAC World Congr.*, Barcelona, Spain, July 2002, CD-ROM Paper 990.

[8] R. S. H. Istepanian, J. F. Whidborne, and P. Bauer, "Stability analysis of block floating point digital controllers," presented at the UKACC Int. Conf. Control, Cambridge, U.K., Sept. 4–7, 2000.

[9] K. F. Man, K. S. Tang, and S. Kwong, *Genetic Algorithms: Concepts and Design*. London, U.K.: Springer-Verlag, 1998.

[10] S. Chen and B. L. Luk, "Adaptive simulated annealing for optimization in signal processing applications," *Signal Processing*, vol. 79, no. 1, pp. 117–128, 1999.

[11] G. S. G. Beveridge and R. S. Schechter, *Optimization: Theory and Practice*. New York: McGraw-Hill, 1970.

[12] A. J. Laub, M. T. Heath, C. C. Paige, and R. C. Ward, "Computation of system balancing transformations and other applications of simultaneous diagonalization reduction algorithms," *IEEE Trans. Automat. Contr.*, vol. AC-32, pp. 115–122, Feb. 1987.

[13] J. Kautsky, N. K. Nichols, and P. Van Dooren, "Robust pole assignment in linear state feedback," *Int. J. Control*, vol. 41, no. 5, pp. 1129–1155, 1985.

[14] G. P. Liu and R. J. Patton, *Eigenstructure Assignment for Control System Design*. Chichester, U.K.: Wiley, 1998.